Journal of Phylogenetics & Evolutionary Biology

Mukherjee and Diehl, J Phylogenetics Evol Biol
2017, 5:1
DOI: 10.4172/2329-9002.1000175

Research Article Open Access

Speciation Genomics of Protein-Coding Genes Common to Mycoplasmatales

Dipaloke Mukherjee[1*] **and Walter J Diehl**[2]

[1]*Department of Food Science, Nutrition and Health Promotion, Mississippi State University, MS 39762, USA*

[2]*Department of Biological Sciences, Mississippi State University, MS 39762, USA*

[*]**Corresponding author:** Dipaloke Mukherjee, Department of Food Science, Nutrition and Health Promotion, Mississippi State University, MS 39762, USA, Tel: +1-662-341-2848; Fax: +1-662-325-8728; E-mail: dipaloke.mukherjee@gmail.com

Received date: Dec 22, 2016; **Accepted date:** Jan 09, 2017; **Published date:** Jan 13, 2017

Abstract

Identifying regions of a genome that evolve by natural selection, particularly as species diverge, has been a matter of considerable interest. The genomes of 12 species in the eubacterial order *Mycoplasmatales* were compared to test the hypothesis that natural selection targets genes by function and/or at given moments in the phylogenetic history of the species. These species possess some of the smallest genomes known, and analyses on the set of protein-coding genes common to all species in the study will shed light on the evolution of some of the most critical genes to living organisms. Genes that control cellular processes showed greater evidence of natural selection than genes of unknown function or genes associated with information processing and storage or metabolism. Moreover evidence of natural selection was only detected in the deepest branches of the *Mycoplasmatales* phylogeny, including one node where a host shift from land plants to insects likely occurred and another node where a host shift from land plants/insects to land vertebrates likely occurred. Many of the genes that showed the strongest evidence of natural selection (e.g. *secA, secY, ftsH, ftsY, yidC, lepA, dnaK*) encode proteins that are components of the Sec-dependent secretory pathway, which regulates the extracellular translocation of proteins. The Sec-dependent secretory pathway is proposed to play a role in speciation of *Mycoplasmatales* by altering the type and amount of secreted proteins, thereby affecting virulence of *Mycoplasma sp*. in response to infection of novel hosts.

Keywords: Minimal genome; *Mycoplasma*; Speciation; Natural selection

Introduction

The proliferation of sequenced genomes has permitted the evaluation of the role of natural selection at that level of organization. Studies have shown that some species, such as Drosophila melanogaster, show positive Darwinian selection in relatively large number of genes [1,2], whereas other species, such as Arabidopsis thaliana show very low levels of positive selection compared to purifying or negative selection [3,4]. By comparison Homo sapiens show intermediate levels of positive selection [5,6]. Such studies have in common an evaluation of selection within a single lineage but generally do not address selection that may be occurring as new species arise, that is at splits in their respective phylogenetic trees. A few studies have targeted selection as species diverge [7,8], but most of them are restricted to evaluating SNPs scattered throughout the genome [9,10] and not whole sequences of genes. Using this approach, one may infer whether selection is common or rare during speciation but not necessarily whether selection is associated with particular functional groups of genes, which in turn may inform hypotheses on the genetics of speciation.

The genus Mycoplasma (Order *Mycoplasmatales*, Domain Eubacteria) is a polyphyletic group [11] that comprises single cell, gram-positive-like, obligate parasites that may cause respiratory, urogenital and other diseases in vertebrates including humans [12,13]. They lack cell walls due to their inability to synthesize peptidoglycan, and they are considered to be the simplest form of self-replicating biological systems but are entirely dependent on the host cells for essential nutrients. Moreover they possess extremely small genomes (range: 524-1053 genes) [12], which make them ideal for studying natural selection at the genome level. As such, the set of protein-coding genes common to all *Mycoplasmatales* species should comprise a set of genes that is among the most functionally critical to living organisms and that has potential for the greatest consequence if selection acts commonly and consistently thereon.

The objectives of this study were to test the hypotheses (1) that natural selection has acted on the same sets of genes at different times in the *Mycoplasmatales* phylogeny, (2) that mutation saturation has not affected the pattern of selection acting on the genes in the Order *Mycoplasmatales,* and (3) that the likelihood of natural selection targeting a particular set of genes has depended on the function of the gene products. This approach has the potential to identify genes (or combinations of genes, gene complexes, or pathways) that may be involved directly or indirectly in speciation.

Materials and Methods

Phylogenetic tree

A Bayesian phylogenetic tree (Figure 1a) was constructed from the 16S ribosomal DNA (rDNA) sequences from 12 *Mycoplasma* species (Table 1), whose genomes had been sequenced and sufficiently annotated as of April 15, 2008. These sequences were obtained from the National Center for Biotechnology Information (NCBI) database

(http://www.ncbi.nlm.nih.gov/) [14]. The 16S rDNA sequences from *Lactobacillus acidophilus* and *Escherichia coli* were used as the out groups. The sequences were aligned with ClustalX [15,16], and MrBayes version 3.1 [17] was used to construct the tree. A General Time Reversible (GTR) model with Gamma-distributed rates (GTR+γ) was determined to be the best fit model for the run by Modeltest version 3.7 [18] and a likelihood ratio test was used for model selection. A total of 100,000 generations of Markov Chain Monte Carlo (MCMC) simulations were run with the first 2500 generations ignored as burnins, and the consensus tree was selected for further analyses. This tree was found to converge after the run, indicated by a final standard deviation of split frequency of 0.006444. Four nodes (A, B, C and D, Figure 1a) were chosen for selection analyses, since these nodes united two clades each with multiple species/variants. A more detailed phylogenetic tree showing the divergence of the *Mycoplasma/Ureaplasma* group from the *Spiroplasma* and *Mesoplasma/Entomoplasma* groups have been presented in the Figure 1b, with the divergence dates of the latter two groups indicated (extrapolated from Maniloff [11]).

Organism	Accession Number	References
M. agalactiae PG2	CU179680	Sirand-Pugnet et al. [24]
M. capricolum subsp capricolum ATCC 27343	CP000123	Craig Venter Institute [25]
M. gallisepticum str R (low)	AE015450	Papazist et al. [26]
M. genitalium G37	L43967	Fraser et al. [27]
M. hyopneumoniae 232	AE017332	Minion et al. [28]
M. hyopneumoniae 7448	AE017244	Vasconcelos et al. [29]
M. hyopneumoniae J	AE017243	Vasconcelos et al. [29]
M. mobile 163K	AE017308	Jaffe et al. [30]
M. penetrans HF-2	BA000026	Sasaki et al. [31]
M. mycoides subsp mycoides PG1	BX293980	Westberg et al. [32]
M. pneumonia M129	U00089	Himmelreich et al. [33]
M. pulmonis UAB CTIP	AL445566	Chambaud et al. [34]
M. synoviae 53	AE017245	Vasconcelos et al. [29]
U. parvum serovar 3 str. ATCC 700970	AF222894	Glass et al. [35]

Table 1: List of the complete genomes used in the study, obtained from the National Center for Biotechnology information (NCBI) database (http://www.ncbi.nlm.nih.gov/).

Genomic database

A database was constructed that contained the 221 protein coding gene sequences common to all of the 12 species of *Mycoplasma* used in the study (Table 1). Sequences and functions were obtained from National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/) database. The Kyoto Encyclopedia of Genes (KEGG) and Genomics (http://www.genome.jp/kegg/) [36] database was used to detect orthologs for a particular gene. Clusters of Orthologus Groups (COG) [37] were used to segregate these genes into functional groups, namely information storage and processing genes, cellular processes genes and metabolism genes plus a group of poorly characterized genes (Table 2).

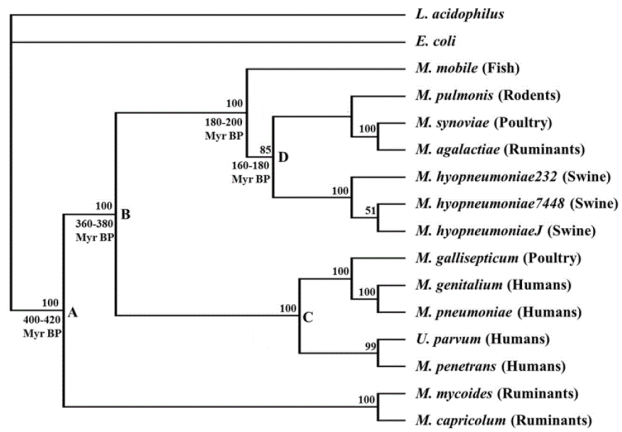

Figure 1a: Bayesian phylogeny from 16S ribosomal DNA (rDNA) sequences for 12 Mycoplasmatales species, using *Lactobacillus acidophilus (L. acidophilus) and Escherichia coli (E. coli)* as outgroups. The clade credibility values for each branch are given. Approximate divergence dates (extrapolated from Maniloff [11]), are indicated for each node. Nodes A, B, C, & D were chosen for tests of natural selection. The specific hosts for each species are indicated in parentheses [12,19-23].

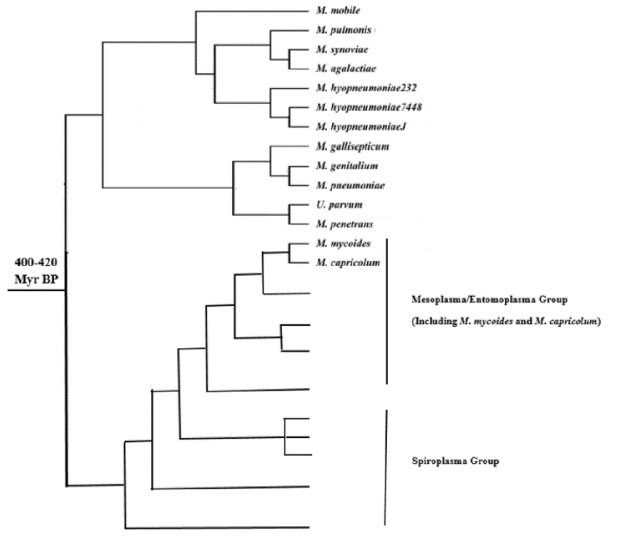

Figure 1b: Mollicutes phylogenetic tree showing the relationships among *Spiroplasma*, *Mesoplasma/Entomoplasma* and *Mycoplasma/Ureaplasma* groups [extrapolated from Maniloff [11] and incorporating phylogeny from the current study].

Functional Category	Number of Genes
Information Processing and Storage Genes	123
Cellular Processes Genes	26
Metabolism Genes	51
Poorly Categorized Genes	21

Table 2: Number of genes belonging to the different functional categories.

Sequence alignment

The software program package Data Analysis in Molecular Biology and Evolution (DAMBE) [38] was used to align the sequences. First, the nucleotide sequences were translated to amino acid sequences. Next, multiple sequence alignments were conducted using ClustalW [16,39] with gap open and extension penalties of 10 and 0.1, respectively and using BLOSUM as the amino acid substitution matrix. Finally, the original nucleotide sequences were realigned against the aligned amino acid sequences. This method is a codon-based approach of sequence alignment, which can prove to be useful in accounting for gaps in the middle of the sequences.

Data analyses

Two separate statistical procedures were followed for the analysis of selection patterns–a codon substitution models test [40] and the McDonald-Kreitman test [41]. The ratio of non-synonymous (dN) to synonymous (dS) nucleotide substitution rates (dN/dS, designated by the letter ω) [42] forms the basis of both tests. A ratio of 1 indicates neutrality. In other words, the rate of fixation of a neutral amino acid mutation will be equivalent to that of a synonymous substitution. A ratio less than 1 indicates purifying selection and the substitution is eventually eliminated from the population. If the ratio is greater than 1, then positive Darwinian selection will retain the amino acid mutation in the population [43]. Codon substitution models tests can be used to compare dN/dS ratios among branches on a phylogenetic tree, where a significant difference implies selection. The McDonald-Kreitman test effectively partitions dN/dS ratios into fixed differences between clades vs. polymorphism across clades, where a significant difference implies selection by either adaptive fixation or polymorphism excess. Because each test uses dN/dS ratios differently, the results are effectively independent.

Codon substitution models tests

The Codeml program of the software program package Phylogenetic Analysis using Maximum Parsimony (PAML) version 4.0 [40] was used to conduct these analyses. First, each gene in the original dataset was analyzed by the free and one ratio models test [40]. The free ratio model assumes different dN/dS ratios for the different branches of the phylogenetic tree, whereas the one ratio model assumes a single dN/dS value for the entire tree. This test effectively determines whether selection at a gene may be occurring somewhere in the phylogeny but it cannot determine where. Log-likelihood values for each of these two models were computed and twice the log-likelihood differences for each of the genes were compared to a $\chi 2$ distribution with $\alpha=0.000057$ (the same α value that was used for the McDonald-Kreitman test after applying a Bonferroni correction-described later) and 24 degrees of freedom (25 branches analyzed under the free ratio model plus one under the one ratio model minus two) [44]. Result from the Fisher's Exact test (P=0.0035) [45,46] showed that the ratio of the genes showing evidence of selection to the ones showing neutrality among functional groups was significantly different.

To determine whether the dN/dS ratios of the different clades in the phylogeney were different from each other as well as from the background, two branch models tests (three vs. two branch ratio models test) [40] were used. The three branch ratio model assumes different dN/dS values for the two clades as well as the background. The two branch ratio model, on the other hand, assumes an equal dN/dS ratio for the two clades that are being compared, while the dN/dS value for the rest of the tree (the 'background') is assumed to be free. These two tests were applied only to the 153 genes that gave evidence of selection in response to the application of free and one ratio models tests. Four clades of the phylogenetic tree with sufficient species/variants to show sequence variation were analyzed this way. Log-likelihood values for these models were computed, and twice the log-likelihood differences were compared to a $\chi 2$ distribution with $\alpha=0.05$ with 3 degrees of freedom (three clades tested under the three branch ratio model plus two under the two branch ratio model minus 2). A three way contingency table (4 functional categories × 4 clades × 2 possible selection results, namely selection or neutrality, Table 3) was constructed to summarize the results from the analyses using the various codon substitution models. Log-linear analysis ($\alpha=0.05$) [47] was conducted on the data to determine whether selection pattern depended on gene function and/or on nodes of the phylogenetic tree (Table 4).

McDonald-Kreitman tests

The software program package DNA Sequence Polymorphism (DnaSP) version 4.20.2 [48] was used to compute the neutrality indices [49] for each gene for each pair of clades tested in the phylogeny. Two-tailed Fisher's exact tests [45] were conducted to assess the significances of the computed NIs, with an α value of 0.000057 (after applying a Bonferroni correction [44], α=[0.05/Number of genes studied, 221 × Number of clades analyzed 4]). These tests were conducted using the software program package DnaSP [48]. Finally, a three way contingency table (4 functional categories × 4 clades × 2 McDonald-Kreitman test results, namely adaptive fixation or neutrality, since none of the genes exhibited polymorphism excess, Table 3) was constructed to summarize the outcomes from the McDonald-Kreitman tests. The results were tested by log-linear analysis ($\alpha=0.05$) [47] to determine whether selection pattern varies depending on functional categories of the gene and/or clades of the phylogenetic tree (Table 4). For both the codon substitution models and the McDonald-Kreitman tests, the VassarStats Web Site for Statistical Computation (http://dogsbody.psych.mun.ca/VassarStats/abc.html; [50]) was used for conducting the log-linear analyses.

Detection of mutation saturation

Mutation saturation [51,52] can be detected by analyzing the frequency of the complex codons in a gene. Complex codons are a group of highly variable codons for which the pattern of non-synonymous and synonymous substitutions for fixed differences and polymorphisms among species sets cannot be inferred, as defined by Rozas et al. [48], and that subsequently cannot be analyzed by the package DnaSP. We assumed that the greater the degree to which a gene shows mutation saturation, the greater the number of complex codons that gene would possess. Thus mutation saturation can be

analyzed by plotting number of complex codons as a function of the total number of codons in a gene for the all genes in the genome. Genes at each node were analyzed separately. Complete mutation saturation would be indicated if all codons were complex. Complete absence of mutation saturation would be indicated if no codons in a gene were complex.

Pathway analysis

The network of the protein-protein interactions of the cellular processes gene products from *M. pneumonia* M129 [33], based on their involvement in interconnected biochemical pathways, 173 was generated in the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) database [53] of known and predicted protein-protein interactions. The software program Cytoscape version 2.8.2 [54] was used to visualize the network.

Results

An initial analysis of natural selection among all species in the study (free vs. one ratio codon substitution models test) [40] indicated that 153 of 221 genes showed preliminary evidence of natural selection (α=0.000057) after applying a Bonferroni correction [44] somewhere in the *Mycoplasmatales* phylogeny. There was a significant difference in the ratio of genes showing selection to genes showing neutrality among functional groups (Fisher exact test, P=0.0035 [45,46]) with 92% of cellular processes genes showing evidence of selection compared to 66% of genes showing selection for other groups combined (Table 3A). The significant results here justified subsequent node by node analyses. Genes that showed neutrality in this initial evaluation were assumed to be neutral in the all subsequent tests.

Nodes A-D (Figure 1a) was chosen for subsequent analyses because the entire respective sister clades nested within them contained multiple species or variants with sequenced genomes, hence the potential for genetic variation at every gene. Of the 153 genes showing evidence of selection in the initial test, 135 genes showed evidence of selection in subsequent three vs. two ratio codon substitution models tests [40] conducted node by node (Table 3B), and 90 genes showed evidence of adaptive fixation (similar to divergent selection) in McDonald-Kreitman tests [41] conducted node by node (Table 3C). In the latter tests, no genes showed a significant excess of polymorhisms.

A	**Codon Substitution Models Tests (free- vs. one-ratio model)**									
Node	**Information Processing and Sotrage Genes**		**Cellular Processes Genes**		**Metabolism Genes**		**Poorly Characterized Genes**		**Total**	
	S	N	S	N	S	N	S	N	S	N
All	74	49	24	2	39	12	16	5	153	68
B	**Codon Substitution Models Tests (three- vs. two-ratio model)**									
Node	**Information Processing and Storage Genes**		**Cellular Processes Genes**		**Metabolism Genes**		**Poorly Characterized Genes**		**Total**	
	S	N	S	N	S	N	S	N	S	N
A	60	63	20	6	31	20	10	11	121	100
B	26	97	5	21	9	42	6	15	46	175
C	0	123	0	26	2	49	0	21	2	219
D	19	104	0	26	4	47	3	18	26	195
Total	105	387	25	79	46	158	19	65	195	689
C	**McDonald-Kreitman Tests**									
Node	**Information Processing and Sotrage Genes**		**Cellular Processes Genes**		**Metabolism Genes**		**Poorly Characterized Genes**		**Total**	
	S	N	S	N	S	N	S	N	S	N
A	30	93	11	15	13	38	7	14	61	160
B	40	83	13	13	10	41	4	17	67	154
C	12	111	7	19	2	49	4	17	25	196
D	7	116	0	26	7	44	0	21	14	207
Total	89	403	31	73	32	172	15	69	167	717

Table 3: Contingency tables of genes showing natural selection (S) or neutrality (N) partitioned among functional groups and nodes in the Mycoplasmatales phylogeny. All genes showing selection were significant in respective tests at α=0.000057 after applying Bonferroni correction

[44]. Cells with ratios of genes showing selection to genes showing neutrality that are arbitrarily 2.5× greater than the respective total S/N ratio are enclosed in boxes; cells with S/N ratios that are 5× greater than the total S/N ratio are enclosed in double boxes. (A) Distribution of genes among functional groups for the entire phylogeny from the free vs. one ratio model of codon substitutional models tests [40]. (B) Distribution of genes among functional groups and nodes from the three vs. two ratio model of codon substitutional models tests [40]. (C) Distribution of genes among functional groups and nodes from McDonald-Kreitman tests [41].

Only 62 genes showed evidence of selection by both codon substitution models tests and McDonald-Kreitman tests. When selection state was partitioned among functional groups and phylogenetic nodes, log-linear analyses [47] showed a significant three-way interaction (P<0.0001) regardless of the selection test used (Table 4), indicating that the pattern of selection depended on both gene function and phylogeny. Moreover, partial interactions of selection state and functional group (P<0.017) and selection state and phylogenetic node (P<0.0001) were also significant (Table 4). Cellular processes genes showed greater evidence of selection (24-30%) than genes in all other functional categories combined (17-22%). Nodes A and B had a greater proportion of genes showing evidence of selection (29-38%) than nodes C and D (6-9%). Only 45 genes showed evidence of selection by both tests at either node A or node B, and 29% of these genes were in the cellular processes category despite accounting for only 12% of genes overall. Regardless of selection test, the greatest proportion of genes showing evidence of selection occurred in the cellular processes category at either node A or B (Table 3).

Three- vs. Two-ratio Codon Substitution Models Test	
Source	**P-values**
Selection State × Gene Function × Node (Phylogeny)	<0.0001***
Selection State × Gene Function (Effect of Node removed)	<0.0177*
Selection State × Node (Effect of Gene Function removed)	<0.0001***
B. McDonald-Kreitman Test	
Selection State × Gene Functional × Node (Phylogeny)	<0.0001***
Selection State × Gene Function (Effect of Node removed)	<0.0012**
Selection State × Node (Effect of Gene Function removed)	<0.0001***

Table 4: Log-linear analysis [47] of association among selection state, gene function, and node in the Mycoplasmatales phylogeny from 3-way contingency tables (Tables 3B and 3C) for (A) codon substitution models tests [40] and (B) McDonald-Kreitman tests [41].

Figures 2a and 2b show the relationships between the total number of codons in a gene and the number of complex codons, indicating mutation saturation. As one would expect, the more codons that a gene possesses, the greater the number of complex codons that occur as well (Table 5). It is evident from the Figures 2a and 2b that in the course of evolution, the genomes of the Mycoplasmatales species have acquired some degree of mutation saturation, especially at the deeper nodes A and B, although in all cases the level is less than required to show complete saturation. But because nodes A and B showed the greatest evidence of selection, there is possibility that this pattern could have been caused by mutation saturation. If that were the case, one would expect that the majority of genes showing selection would occur above or below the regression line, which did not occur at any node. That is at all nodes, the genes showing natural selection showed no more tendencies toward mutation saturation than genes showing neutrality (Figure 2a). Further, if evolutionary rate varies with gene function, then mutation saturation could theoretically, albeit not necessarily, lead to an apparent excess of natural selection in functional groups with greater evolutionary rates. If this were the case, the slopes of the relationship between the number of complex codons and the total number of codons in a gene would be greater for functional groups showing excess natural selection. Not only is this not the case at any node, but also the slope of the aforementioned relationship for cellular processes genes is less than that for all functional groups combined at each node (Table 5).

Node	**Functional Category**	**Slope**	**r^2**	**p-Value**
A	Information Processing and storage	0.5428	0.8605	<0.000*
	Cellular Processes	0.4654	0.6523	<0.000*
	Metabolism	0.2876	0.4422	<0.000*
	Poorly Characterized	0.4507	0.7776	<0.000*
	All Functional Categories combined	0.4968	0.7836	<0.000*
B	Information Processing and storage	0.4781	0.8417	<0.000*
	Cellular Processes	0.3793	0.6027	<0.000*
	Metabolism	0.2699	0.4549	<0.000*
	Poorly Characterized	0.3876	0.7802	<0.000*
	All Functional Categories combined	0.4357	0.767	<0.000*
C	Information Processing and storage	0.1489	0.7273	<0.000*
	Cellular Processes	0.0826	0.3055	<0.000*
	Metabolism	0.0785	0.2361	0.0003*
	Poorly Characterized	0.0643	0.3732	0.0033*
	All Functional Categories combined	0.1281	0.5912	<0.000*
D	Information Processing and storage	0.1675	0.7319	<0.000*
	Cellular Processes	0.1552	0.4914	<0.000*
	Metabolism	0.1048	0.3265	<0.000*
	Poorly Characterized	0.1422	0.7046	<0.000*
	All Functional Categories combined	0.158	0.6517	<0.000*

Table 5: The r^2 and p-values for the relationship between the total number of codons and the number of complex codons in a gene for all genes at each node. The p-values that were significant at α=0.0125 [after applying Bonferroni correction (0.05/4, the number of nodes tested)] are indicated with an asterisk (*).

Using *M. pneumoniae* M129 [33] as a reference species, pathway analysis based on the protein-protein interaction patterns of the total set of proteins encoded by cellular process genes revealed some evidence of clustering within two subdivisions, namely within post-translational modification, protein turnover and chaperone proteins, and within inorganic ion transport and metabolism proteins (Figure 3). Proteins with the greatest number of interactions (*secA, ffh, secY,*

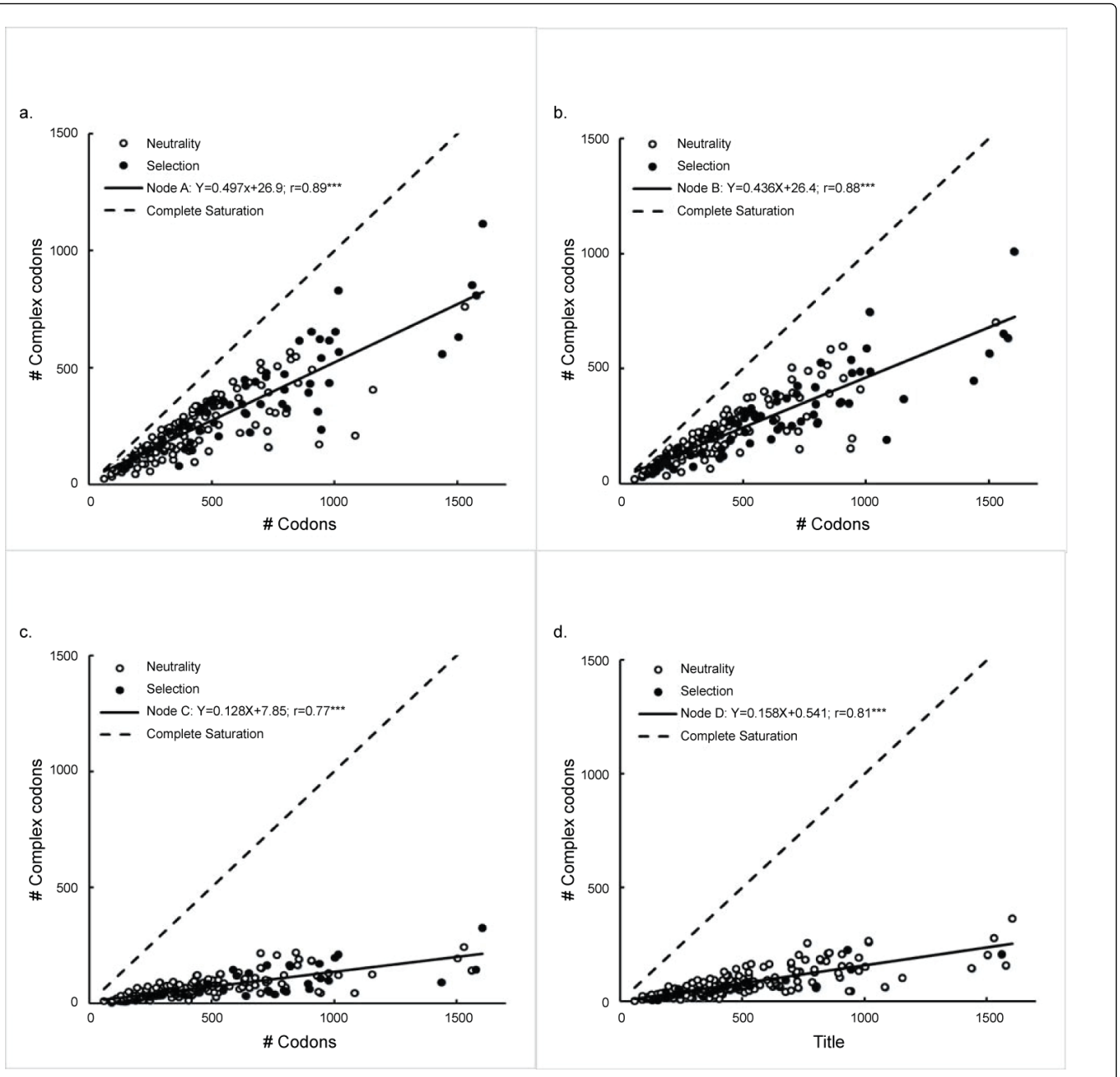

Figure 2a: Relationship between the total number of codons and number of complex codons [48] (solid lines from linear regressions) for genes showing selection under the McDonald and Kreitman tests [41] (closed circles) and genes showing neutrality (open circles) at nodes A (a), B (b), C (c), and D (d) in the phylogeny of the Mycoplasmatales. Complete saturation would be indicated by the dashed lines.

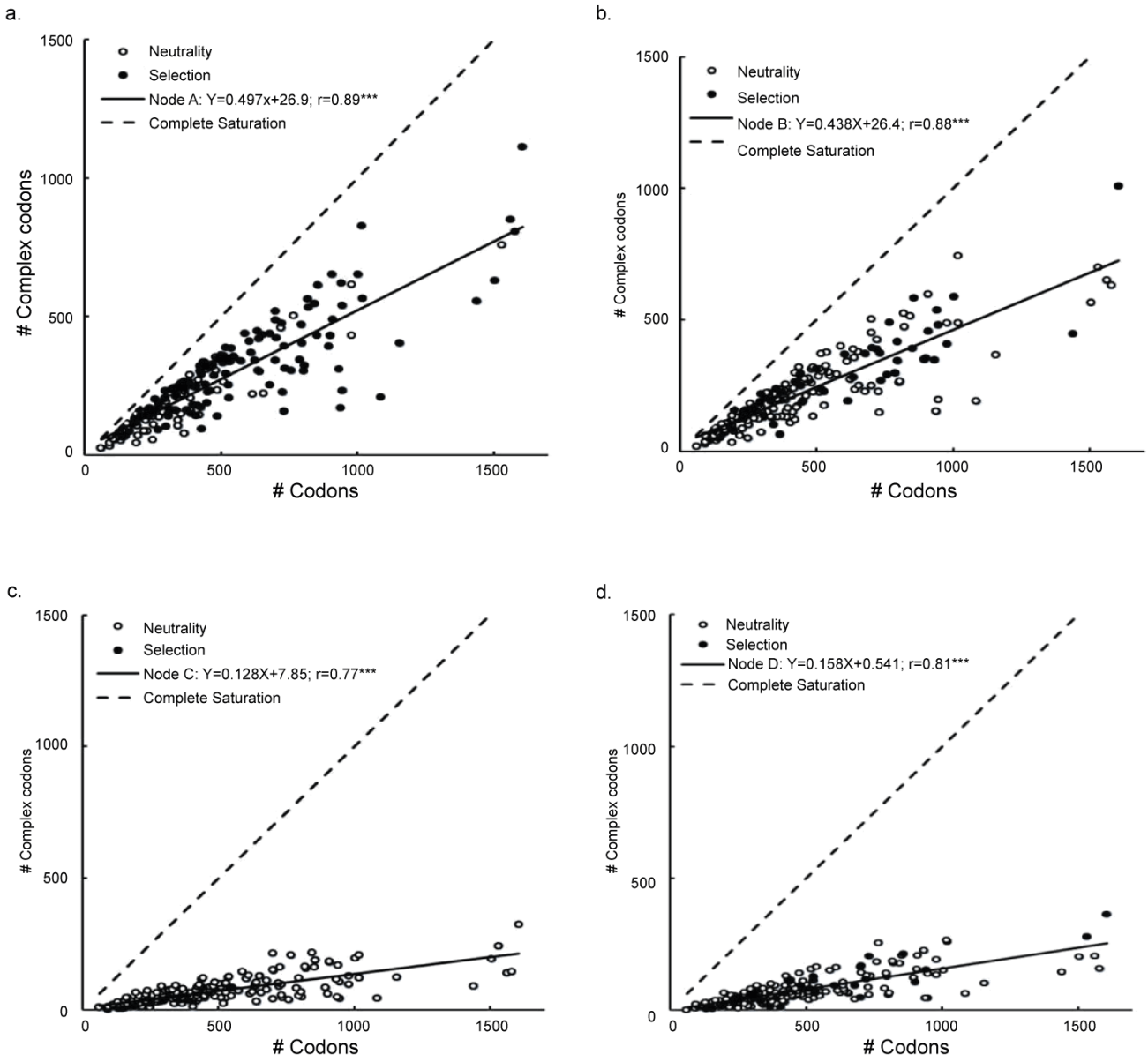

Figure 2b: Relationship between the total number of codons and number of complex codons [48] (solid lines from linear regressions) for genes showing selection under the codon substitution models tests, Yang [40] (closed circles) and genes showing neutrality (open circles) at nodes A (a), B (b), C (c), and D (d) in the phylogeny of the Mycoplasmatales. Complete saturation would be indicated by the dashed lines.

dnaK and lepA) are distributed among three functional subdivisions but are all components or putative components of the *Mycoplasmatales* Sec-dependent secretory pathway [55]. Moreover, a significantly greater proportion of genes encoding these proteins (7 of 8) showed natural selection at both nodes A and B compared to that for genes encoding proteins outside the Sec-dependent secretory pathway (6 of 18; Fisher Exact Test, $P<0.05$).

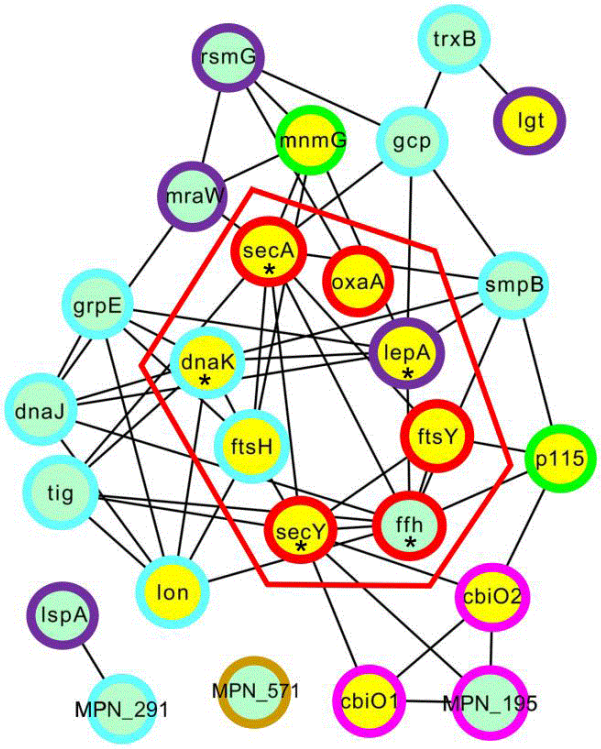

Figure 3: Interaction patterns among the proteins based on involvement in interconnected biochemical pathways encoded by cellular processes genes from *M. pneumoniae* M129 [33]. Proteins showing evidence of natural selection at both Nodes A and B are indicated with yellow inner circles; otherwise inner circles are green. Proteins in the Sec-dependent secretory pathway are bounded by the red polygon. Proteins showing 7 or more interactions are indicated with asterisks. Post translational modification, protein turnover and chaperone proteins (blue outer ring): molecular chaperon DnaK (dnaK), ATP dependent heat shock protease Lon (lon), cell division protein FtsH (ftsH), heat shock protein GrpE (grpE), SSRA-binding protein (smpB), trigger factor (tig), o-sialoglycoprotein endopeptidase (gcp), molecular chaperone/heat shock protein DnaJ (dnaJ), thioredoxin reductase (trxB), glycoprotease family protein (MPN_291). Intracellular trafficking and secretion proteins (red outer ring): cell division protein FtsY (ftsY), preprotein translocase subunit SecA (secA), preprotein translocase subunit SecY (secY), inner membrane protein translocase YidC (oxaA), signal recognition particle protein/GTPase (ffh). Cell wall/membrane biogenesis proteins (purple outer ring): prolipoprotein diacylglyceryl transferase (lgt), GTP-binding protein LepA (lepA), glucose-inhibited division protein B/methyltransferase GidB (rsmG), signal peptidase II (lspA), s-adenosyl-methyltransferase (mraW). Inorganic ion transport and metabolism proteins (pink outer ring): cobalt transporter ATP binding subunit (cbi01), cobalt transporter ATP binding subunit (cbi02), cobalt ABC transporter permease protein (MPN_195). Cell cycle control, mitosis and meiosis proteins (green outer ring): chromosomal segregation protein SMC (p115), tRNA-uracil-5-carboxymethylaminomethyl modification enzyme (mnmG). Defense mechanisms protein (brown outer ring): ABC transporter ATP-binding and permease protein (MPN_571).

Discussion

The first objective of this study was to test the hypotheses that natural selection has acted on the same sets of genes across different branches in the *Mycoplasmatales* phylogeny, with greater number of genes exhibiting selection at the deeper nodes of the phylogeny. The excess of cellular process genes showing evidence of selection at both nodes A and B indicates the possibility of an evolutionary process that was influenced by selection acting on a common set of genes as species, represented by contemporary clades, diverged early in *Mycoplasmatales* evolution. Failure to detect sufficient evidence of natural selection by both the statistical tests in more recent nodes C and D may indicate (1) that selection was acting on a suite of genes that is unique to a particular clade or species and that therefore was not evaluated here, (2) that selection was acting on regulatory rather than structural regions of the genome (because regulatory sequences do not encode proteins), the statistical methods used cannot assess the pattern of natural selection acting on them, (3) that evolutionary divergence was dominated by neutral processes, an unlikely scenario given evidence from nodes A & B, or (4) that there has not been enough time for non-synonymous substitutions to accumulate sufficiently to reveal evidence of natural selection, another unlikely since divergence at nodes C–D was preceded by a period of rapid genomic change in *Mycoplasma* groups about 190 Myr ago [11]. Regardless, in spite of the inability to detect selection in recent nodes, the study proposes a common pattern of selection acting at more primitive nodes.

The genes showing evidence of natural selection were detected mostly in the deepest nodes of the *Mycoplasmatales* phylogeny. It can be speculated by evaluation of this region of the *Mycoplasmatales* phylogeny (Figure 1a) that early species divergence coincided with the origin of insects (396-407 Myr BP) [56] during the early Devonian (node A) and with the origin of land vertebrates (Amphibians) (368 Myr BP) [57] during the late Devonian (node B). This is consistent with the fact that the *M. mycoides/capricolum* clade comprises the Entoplasmatales group within which *M. mycoides* and *M. capricolum* are derived species [11] that have likely infected vertebrate hosts independently of that by other *Mycoplasmatales* and are thus taxonomically grouped in the genus Mycoplasma by convergence. The *Entoplasmatales* includes the basal group *Spiroplasma* [11], whose hosts are plants and insects [12]. The data suggest that early in the evolution of *Mycoplasmatales* natural selection promoted speciation in response to novel environments associated with host shifts as plants, insects, and vertebrates' successively colonized land and were infected by species of Mycoplasma. Speciation under such circumstances is not unexpected. For example, it has been reported that especially repeat-rich parts of the genome of different lineages of the Irish potato famine pathogen *Phytophthora infestans* evolve through host jumps [58].

The second objective was to test the hypotheses that mutation saturation [51,52] has not affected by causing a false-positive pattern of selection acting on the genes in the order *Mycoplasmatales*. Mutation saturation is a neutral genetic phenomenon that can potentially cause a significant but spurious pattern of selection. The process occurs when a particular base mutates to a different one, then mutates back to the original state-for example, an A mutating to a T, then back to A, or when multiple mutations occur at a given site (an A mutating to G, then to T). In such a case, it is difficult to determine if a particular base underwent two successive mutations in its evolutionary history or none, confounding one's ability to assess the number of non-synonymous and synonymous changes. If mutation saturation is more

likely to occur or persist in one type of change than the other, then selection may be inappropriately implicated. The members of the order *Mycoplasmatales* have high rates of mutation [12], which is one of the primary driving forces behind their evolution, and thus mutation saturation is likely. Indeed our analyses showed that the Mycoplasma genomes have acquired some mutation saturation, especially at the deeper nodes. However at each node, the distribution of genes showing selection vs. genes showing neutrality are essentially the same with regard to mutation saturation, indicating that mutation saturation is not responsible for the patterns of selection observed. Further, it is not at all likely that mutation saturation could produce a pattern of selection that varies with gene function as is seen in this study, since excess mutation saturation, where it existed, tended to occur in functional groups not showing high levels of natural selection.

The final objective was to test the hypotheses that the likelihood of natural selection targeting a particular set of genes in the order *Mycoplasmatales* has depended on the function of the gene products, which has finally influenced speciation. In this study, natural selection has been shown to target cellular processes genes in general and Sec-dependent secretory pathway genes in particular. The Sec-dependent secretory pathway is ubiquitous to living organisms [59] and in the *Mycoplasmatales* functions in extracellular protein transport [59], including the export of proteins affecting virulence–a scenario that was hypothesized to play a role in the evolution of the *Phytoplasmas*, the plant pathogenic group within the class *Mollicutes* [60]. It has been indicated that genes for protein export should be part of the predicted minimal genome in bacteria [61]. We hypothesize that natural selection, acting on genes encoding proteins of the Sec-dependent secretion pathway, has caused speciation by altering the type and amount of secreted proteins, thereby affecting virulence of the *Mycoplasmatales* in response to infection of novel hosts.

The current approach to identifying genes that may play a role in a previous speciation event is very conservative as the final set of genes that has passed through 6 filters (3 natural selection tests, 2 selection-by-function interaction tests, and 1 association by specific common function analysis). It provides a mechanism for identifying the signature of an original selection event that may have been buried in subsequent accumulated genetic variation and sweeps that fixed differences in many other genes.

References

1. Sawyer SA, Kulathinal RJ, Bustamante CD, Hartl DL (2003) Bayesian analysis suggests that most amino acid replacements in *Drosophila* are driven by positive selection. J Mol Evol 57: S154-S164.
2. Sella G, Petrov DA, Przeworski M, Andolfatto P (2009) Pervasive natural selection in the *Drosophila* genome? PLoS Genet 5: e1000495.
3. Bustamante CD, Nielsen R, Sawyer SA, Olsen KM, Purugganan M, et al. (2002) The cost of inbreeding in Arabidopsis. Nature 416: 531-534.
4. Wright SI, Gaut BS (2005) Molecular population genetics and the search for adaptive evolution in plants. Mol Biol Evol 22: 506-519.
5. Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, et al. (2005) Natural selection on protein-coding genes in the human genome. Nature 437: 1153-1157.
6. Coop G, Pickrell JK, Novembre J, Kudaravalli S, Li J, et al. (2009) The role of geography in human adaptation. PloS Genet 5: e1000500.
7. Ahmed M, Liang P (2013) Study of modern human evolution via comparative analysis with the Neanderthal genome. Genomics Inform 11: 230-238.
8. Crisci JL, Wong A, Good JM, Jensen JD (2011) On characterizing adaptive events unique to modern humans. Genome Biol Evol 3: 791-798.
9. Akey JM, Zhang G, Zhang K, Jin L, Shriver MD (2002) Interrogating a high-density SNP map For signatures of natural selection. Genome Res 12: 1805-1814.
10. Neafsey DE, Schaffner SF, Volkman SK, Park D, Montgomery P, et al. (2008) Genome-wide SNP genotyping highlights the role of natural selection in *Plasmodium falciparum* population divergence. Genome Biol 9: R17.
11. Maniloff J (2002) Phylogeny and evolution. In Molecular biology and pathogenicity of mycoplasmas. In: Razin S and Herrmann R (eds.) Kluwer Academic/Plenum Publishers, NY, USA, pp: 31-44.
12. Razin S, Yogev D, Naot Y (1998) Molecular biology and pathogenicity of Mycoplasmas. Microbiol Mol Biol Rev 62: 1094-1156.
13. Belloy L, Janovsky M, Vilei EM, Pilo P, Giacometti M, et al. (2003) Molecular epidemiology of Mycoplasma conjunctivae in caprinae: transmission across species in natural outbreaks. Appl Environ Microbiol 69: 1913-1919.
14. http://www.ncbi.nlm.nih.gov/
15. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25: 4876-4882.
16. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, et al. (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23: 2947-2948.
17. Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic tree. Bioinformatics 17: 754-755.
18. Posada D, Crandall KA (1998) MODELTEST: testing the model of DNA substitution. Bioinformatics 14: 817-818.
19. Al-Momani W, Nicholas RA, Janakat S, Abu-Basha E, Ayling RD (2006) The *in vitro* effect of six antimicrobials against *Mycoplasma putrefaciens*, *Mycoplasma mycoides* subsp. mycoides LC and *Mycoplasma capricolum* subsp. capricolum isolated from sheep and goats in Jordan. Trop Anim Health Prod 38: 1-7.
20. Bergonier D, Berthelot X, Poumarat F (1997) Contagious agalactia of small ruminants: current knowledge concerning epidemiology, diagnosis and control. Rev Sci Tech 16: 843-873.
21. Noormohammadi AH (2007) Role of phenotypic diversity in pathogenesis of avian mycoplasmosis. Avian Pathol 36: 439-444.
22. Opriessnig T, Thacker EL, Yu S, Fenaux M, Meng XL, et al. (2004) Experimental reproduction of postweaning multisystemic wasting syndrome in pigs by dual infection with *Mycoplasma hyopneumoniae* and porcine circovirus type 2. Vet Pathol 41: 624-640.
23. Stadtländer C, Kirchhoff H (1990) Surface parasitism of the fish mycoplasma Mycoplasma mobile 163 K on tracheal epithelial cells. Vet Microbiol 21: 339-343.
24. Sirand-Prugner P, Lartigue C, Merenda M, Jocob D, Barre A, et al. (2007) Being pathogenic, plastic, and sexual while living with a nearly minimal bacterial genome. Plos Genet 3: e75.
25. https://www.crunchbase.com/organization/j-craig-venter-institute#/entity
26. Papazist L, Gorton TS, Kulish G, Markham PF, Browning GF et al. (2003) The complete genome sequence of the avian pathogen *Mycoplasma gallisepticum* strain R (low). Microbiol 149: 2307-2316.
27. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, et al. (1995) The minimal gene complement of *Mycoplasma genitalium*. Science 270: 397-403.
28. Minion FC, Lefkowitz EJ, Madsen ML, Cleary BJ, Swartzell SM, et al. (2004) The genome sequence of *Mycoplasma hyopneumoniae* strain 232, the agent of swine mycoplasmosis. J Bacteriol 186: 7123-7133.
29. Vasconcelos AT, Ferreira HB, Bizarro CV, Carvalho MO, Pinto PM et al. (2005) Swine and poultry pathogens: the complete genome sequences of two strains of *Mycoplasma hyopneumoniae* and a strain of *Mycoplasma synoviae*. J Bacteriol 187: 5568-5577.

30. Jaffe JD, Stange-Thomann N, Smith C, DeCaprio D, Fisher S, et al. (2005) The complete genome and proteome of Mycoplasma mobile. Genome Res 14: 1447-1461.
31. Sasaki Y, Ishikawa J, Yamashita A, Oshima K, Kenri T, et al. (2002) The complete genomic sequence of *Mycoplasma penetrans*, an intracellular bacterial pathogen in humans. Nucleic Acids Res 30: 5293-5300.
32. Westberg J, Persson A, Holmberg A, Goesmann A, Lundeberg J, et al. (2004) The Genome Sequence of *Mycoplasma mycoides* subsp. mycoides SC type strain PG1T, the causative agent of contagious bovine pleuropneumonia (CBPP). Genome Res 14: 221-227.
33. Himmelreich R, Hilbert H, Plagens H, Pirkl E, Li BC, et al. (1996) Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumonia*. Nucleic Acids Res 24: 4420-4449.
34. Chambaud I, Hellig R, Ferns S, Samson D, Galisson F, et al. (2001) The complete genome sequence of the murine respiratory pathogen *Mycoplasma pulmonis*. Nucleic Acids Res 29: 2145–2153.
35. Glass JI, Lefkowitz EJ, Glass JS, Heiner CR, Chen EY, et al. (2000) The complete sequence of the mucosal pathogen Ureaplasma urealyticum. Nature 407: 757-762.
36. www.genome.jp/kegg/
37. Tatusov RL, Galperin LY, Natale DA, Koonin EV (2000) The COG database: a tool for a tool for genome scale analysis of protein functions and evolution. Nucleic Acids Res 28: 33-36.
38. Xia X, Xie Z (2001) DAMBE: Software package for data analysis in molecular biology and evolution. J Hered 4: 371-373.
39. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673-4680.
40. Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 3: 555-556.
41. McDonald JH, Kreitman M (1991) Adaptive evolution at the Adh Locus in *Drosophila*. Nature 351: 652-654.
42. Miyata T, Yasunaga T (1980) Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. J Mol Evol 16: 23-26.
43. Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitution. Mol Biol Evol 3: 418-426.
44. Cabin RJ, Mitchell RJ (2000) To Bonferroni or not to Bonferroni: when and how are the questions. ESA Bull 81: 246-248.
45. Fisher RA (1922) On the interpretation of $\chi 2$ from contingency tables and the calculation of p. J R Stat Soc 85: 87-94.
46. Statistical Applicatory System (SAS) software version 9.1.3, SAS System for Microsoft® Windows® Copyright © 2009, SAS Institute Inc., Cary, NC, USA.
47. King RJ, Plosser CI, Rebelo ST (1988) Production, growth and business cycles: 1. The basic neoclassical model. J Monet Econ 31: 195-232.
48. Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R (2003) DnaSP: DNA polymorphism analysis by the coalescent and other methods. Bioinformatics 19: 2496-2497.
49. Rand DA, Kann LM (1996) Excess amino acid polymorphism among mitochondrial genes from *Drosophila*, mice and humans. Mol Biol Evol 13: 735-748.
50. Lowry R (2010) VassarStats Website for Statistical Computation.
51. Henn BM, Gignoux CR, Feldman MW, Mountain JL (2009) Characterizing the time dependency of human mitochondrial DNA mutation Rate estimates. Mol Biol Evol 26: 217-230.
52. Ho SYW, Phillips, MJ, Cooper A, Drummond AJ (2005) Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. Mol Biol Evol 22: 1561-1568.
53. Von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, et al. (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res 33: D433-D437.
54. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13: 2498-2504.
55. Stephenson K (2005) Sec-dependent protein translocation across biological membranes: evolutionary conservation of an essential protein transport pathway. Mol Membr Biol 22: 17-28.
56. Engel MS, Grimaldi DA (2004) New light shed on the oldest insect. Nature 427: 627-630.
57. Niedźwiedzki G, Szrek P, Narkiewicz K, Narkiewicz M, Ahlberg PE (2010) Tetrapod trackways from the early Middle Devonian period of Poland. Nature 463: 43-48.
58. Raffaele S, Farrer RA, Cano LM, Studholme DJ, MacLean D, et al. (2010) Genome evolution followed host jumps in the Irish potato famine pathogen lineage. Science 330: 1540-1540
59. Staats CC, Boldo J, Broetto L, Vainstein M, Schrank A (2007) Comparative genome analysis of proteases, oligopeptide uptake and secretion systems in *Mycoplasma* spp. Genet Mol Biol 30: 225-229.
60. Bai X, Zhang J, Ewing A, Miller SA, Radek AJ, et al. (2006) Living with genome instability: the adaptation of phytoplasmas to diverse environments of their insect and plant hosts. J Bacteriol 188: 3682-3696.
61. Mushegian AR, Koonin EV (1996) A minimal gene set for cellular life derived by comparison of complete bacterial genomes. Proc Natl Acad Sci 93: 10268-10273.